

EFFECT OF MODEL-DEPENDENT COVARIANCE MATRIX FOR STUDYING BARYON ACOUSTIC OSCILLATIONS

A. LABATIE¹ AND J.L. STARCK

Laboratoire AIM (UMR 7158), CEA/DSM-CNRS-Université Paris Diderot, IRFU, SEDI-SAP, Service d'Astrophysique, Centre de Saclay, F-91191 Gif-Sur-Yvette cedex, France

M. LACHIÈZE-REY

Astroparticule et Cosmologie (APC), CNRS-UMR 7164, Université Paris 7 Denis Diderot, 10, rue Alice Domon et Léonie Duquet F-75205 Paris Cedex 13, France

Draft version November 12, 2012

ABSTRACT

Large-scale structures in the Universe are a powerful tool to test cosmological models and constrain cosmological parameters. A particular feature of interest comes from Baryon Acoustic Oscillations (BAOs), which are sound waves traveling in the hot plasma of the early Universe that stopped at the recombination time. This feature can be observed as a localized bump in the correlation function at the scale of the sound horizon r_s . As such, it provides a standard ruler and a lot of constraining power in the correlation function analysis of galaxy surveys. Moreover the detection of BAOs at the expected scale gives a strong support to cosmological models. Both of these studies (BAO detection and parameter constraints) rely on a statistical modeling of the measured correlation function $\hat{\xi}$. Usually $\hat{\xi}$ is assumed to be gaussian, with a mean ξ_θ depending on the cosmological model and a covariance matrix C generally approximated as a constant (i.e. independent of the model). In this article we study whether a realistic model-dependent C_θ changes the results of cosmological parameter constraints compared to the approximation of a constant covariance matrix C . For this purpose, we use a new procedure to generate lognormal realizations of the Luminous Red Galaxies sample of the Sloan Digital Sky Survey Data Release 7 to obtain a model-dependent C_θ in a reasonable time. The approximation of C_θ as a constant creates small changes in the cosmological parameter constraints on our sample. We quantify this modeling error using a lot of simulations and find that it only has a marginal influence on cosmological parameter constraints for current and next-generation galaxy surveys. It can be approximately taken into account by extending the 1σ intervals by a factor ≈ 1.3 .

Subject headings: large-scale structure of Universe - distance scale - dark energy - cosmological parameters

1. INTRODUCTION

One of the most important question in modern cosmology is to understand the nature of dark energy. This mysterious form of energy is responsible for the accelerate expansion of the Universe, and seems to account for more than 70% of the energy content of the Universe (see e.g. Komatsu et al. (2009); Amanullah et al. (2010); Blake et al. (2011b)).

The acceleration of the expansion of the Universe was first measured with high-redshift supernovae (Riess et al. 1998; Perlmutter et al. 1999). The principle is to use Type Ia supernovae as standard candles in order to probe the redshift-distance relation. The same principle has been used more recently in the study of galaxy clustering at low redshift using Baryon Acoustic Oscillations (BAOs, Bassett & Hlozek (2010)). These structures are remnants of acoustic waves which travelled in the plasma before recombination, when baryons and photons were coupled together. Their absolute size is given by the sound horizon scale at the baryon drag epoch, and is well constrained by measurements of the Cosmic Microwave Background (CMB), $r_s = 153.3 \pm 2$ Mpc (Komatsu et al. 2009). Thus they can be used as a standard ruler to

probe the redshift-distance relation.

BAOs are a very promising cosmological probe because they are less affected by systematics than other methods (Albrecht et al. 2006). They can also be very useful to cross-check results from other probes. This has been done for example in Blake et al. (2011b), where the combination of the WiggleZ, Sloan Digital Sky Survey (SDSS) and 6-degree Field (6dF) surveys have been used to cross-check supernovae results. As future experiments will provide more precise information, it will be critical to correctly analyze and combine these different probes. In particular one might face new challenges to deal with systematic effects that were under statistical uncertainty in previous experiments and that become important.

Possible systematics can come from incorrect statistical modeling of the data. For example in the case of BAOs in large-scale clustering, a classical procedure is to measure the correlation function $\hat{\xi}$ and fit it to an expected correlation function ξ_θ with a dependence on cosmological parameters θ . More precisely, one assumes a statistical model for $\hat{\xi}$ as a function of θ in order to compute the likelihood $\mathcal{L}_\theta(\hat{\xi})$. A common statistical model is to consider that $\hat{\xi}$ is simply gaussian, centered on the expected correlation ξ_θ and with a constant covariance

matrix C (i.e. independent of θ).

The Gaussianity has been shown to be well verified, e.g. in Labatie et al. (2012b) and Manera et al. (2012). However the approximation of a constant covariance C has not been well studied, probably because it is very difficult to estimate a model-dependent covariance matrix C_θ . Indeed the usual procedure to estimate a covariance matrix is to use a large number of realistic mock catalogues and compute the empirical covariance matrix

$$C_{ij} = \frac{1}{N-1} \sum_{k=1}^N [\hat{\xi}_k(r_i) - \bar{\xi}(r_i)][\hat{\xi}_k(r_j) - \bar{\xi}(r_j)] \quad (1)$$

$$\bar{\xi} = \frac{1}{N} \sum_{k=1}^N \hat{\xi}_k \quad (2)$$

Having a good estimate of the covariance matrix requires a lot of simulations. This procedure can already be long for one value of θ , and it seems infeasible to apply it on a multi-dimensional grid of θ values.

As an alternative, one could find analytical formulae to estimate the covariance matrix of the correlation function $\hat{\xi}$. A recent attempt has been made in Xu et al. (2012). It starts from the analytic computation of the covariance matrix of $\hat{\xi}$ for a Gaussian density field. The covariance matrix is further modified to better match the empirical covariance matrix on mock catalogues. It is shown to reproduce the empirical covariance matrix obtained with mock catalogues, while regularizing it.

This is very interesting because it provides with little effort the covariance matrices for different input power spectra $P(k)$ of the galaxy field, i.e. a model-dependent covariance matrix C_θ . However the procedure is not totally blind and requires an ad hoc fitting of the covariance matrix to mock catalogues for a given model. In particular it has not been shown that the resulting model-dependent covariance matrix C_θ is also a good estimate for other models than the one used for the fitting.

In this article we do not study this question of analytically modeling the covariance matrix. Instead we study whether this modeling is actually required, i.e. if the model-dependence of C_θ affects the statistical analysis (e.g. by changing confidence regions). We will restrict to cosmological parameter constraints using the correlation function (we will not look at the question of BAO detection for reasons explained in section 3.2).

For our analysis to be feasible, we will only consider 3 parameters in θ that have the most impact on the expected correlation function ξ_θ . The first parameter is the matter density $\omega_m = \Omega_m h^2$ which determines the horizon scale at the matter-radiation equality ($\propto \omega_m^{-1}$). It also has a little influence on the sound horizon scale ($\propto \omega_m^{-0.25} \omega_b^{-0.08}$ with $\omega_b = \Omega_b h^2$ the baryon density) and changes the amplitude of the BAO peak (for a constant ω_b). The second parameter is α , that determines how the correlation function is dilated when using a fiducial cosmology instead of the true cosmology to convert redshifts into distance. This parameter is the one that really probes the distance-redshift relation and it is mostly constrained by the position of the BAO peak. Finally the third parameter is a constant bias $B = b^2$ in the correlation function that accounts for different amplitude effects (linear redshift distortions, linear galaxy bias, amplitude

of matter fluctuation σ_8).

As we will estimate the covariance matrices using mock catalogues, a parameterization of C_θ with a 3-dimensional parameter $\theta = (\omega_m, \alpha, B)$ may already seem infeasible. However we will show how to optimize our simulations and the computation of the correlation function in order to make it feasible. We will show that there is in fact only 1 parameter that needs to be varied, and that the 2 other parameters can be taken into account without adding much effort.

The plan of this paper is as follows: we start in section 2 by describing the SDSS DR7-Full data catalogue that we use. In section 3 we discuss the correlation function modeling and estimation. Section 4 presents our new procedure to estimate a model-dependent covariance matrix C_θ with a 3-dimensional parameter $\theta = (\omega_m, \alpha, B)$ in a reasonable time. In section 5 we give results on the statistical modeling of the correlation estimator $\hat{\xi}$: absence of bias in $\hat{\xi}$, Gaussianity of $\hat{\xi}$, dependence of the covariance matrix C_θ on ω_m , α and B . Finally in section 6 we study the modeling error in parameter constraints due to the approximation of C_θ as a constant C . We study this modeling error on the SDSS DR7-Full $\hat{\xi}$ and we perform a quantitative analysis using simulations.

2. DATA CATALOGUE

In this study we use the Luminous Red Galaxies sample (LRG) sample of the last Data Release 7 (DR7) of the SDSS. LRGs are selected using the algorithm in Eisenstein et al. (2001) which consists in different luminosity and color cuts using the five passbands u, g, r, i and z . These galaxies are very luminous and good tracers of massive dark matter haloes. The sample is quasi-volume-limited (i.e. nearly of constant density) up to redshift $z \approx 0.36$ and extends up to $z \approx 0.47$ in a flux-limited way. In order to convert redshifts into distances we use a flat Λ CDM fiducial cosmology with $\Omega_m = 0.25$. We plot the resulting density of the catalogue in figure 1.

We use the DR7-Full sample of the analysis in Kazin et al. (2010) that is available online¹ and has the characteristics given in table 1.

TABLE 1

# of LRGs	105,831
z_{\min}	0.16
z_{\max}	0.47
$\langle z \rangle$	0.324
$M_{g,\min}$	-23.2
$M_{g,\max}$	-21.2
$\langle M_g \rangle$	-21.72
Area (deg ²)	7,908
Volume ($h^{-3} \text{Gpc}^3$)	1.58
Density ($10^{-5} h^3 \text{Mpc}^{-3}$)	6.70

NOTES.—Characteristics of the SDSS LRG sample used DR7-Full from Kazin et al. (2010). Volume and density have been computed with a flat Λ CDM fiducial cosmology with $\Omega_m = 0.25$.

The sample is mostly contiguous, with only 9.8% outside of the main part of the Northern Galactic Cap. The

¹ <http://cosmo.nyu.edu/~eak306/SDSS-LRG.html>

number of LRGs is equal to 96763 in the Northern Galactic Cap and 9068 in the Southern Galactic Cap. We show the footprint of the survey in figure 2 with the Northern contiguous part and the few stripes in the Southern part (the blue line represents the Galactic plane).

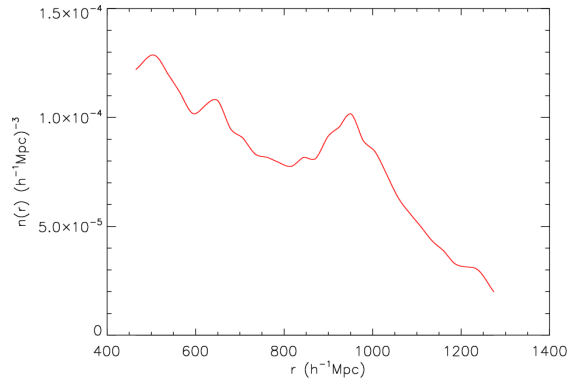


FIG. 1.— Observed density of the sample DR7-Full when using a flat Λ CDM fiducial cosmology with $\Omega_m = 0.25$ to convert redshifts into distances.

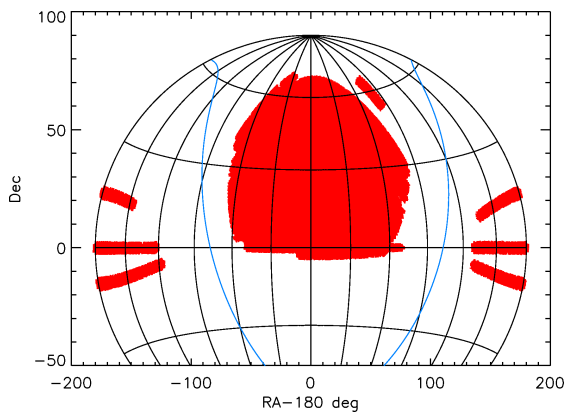


FIG. 2.— SDSS DR7-Full sample sky coverage in Aitoff projection. The solid blue line represents the Galactic plane which separates the Northern Contiguous region and the Southern region.

3. CORRELATION FUNCTION MODELING AND ESTIMATION

3.1. Correlation function modeling

The correlation function is a second order statistic that measures the clustering of a continuous continuous field or a point process. For the galaxy field, it measures the excess of probability to find a pair of galaxies in volumes dV_1 and dV_2 separated by \mathbf{x} compared to a random unclustered distribution

$$dP_{12} = \bar{n}[1 + \xi(\mathbf{x})]dV_1dV_2 \quad (3)$$

with \bar{n} the mean density of points. Due to the cosmological principle the correlation function $\xi(\mathbf{r})$ is isotropic so that it only depends on the norm of the separation vector $r = \|\mathbf{r}\|$. However we do not exactly measure the correlation function for two reasons

- We observe galaxies in redshift space so that there are redshift distortions in the line of sight direction

- The choice of fiducial cosmology dilates the galaxy survey differently in the line of sight and transverse directions

As explained later the second effect can be neglected, i.e. we can model the effect of a wrong fiducial cosmology by a single dilation factor α in all directions. We still want to measure the correlation function as a function of $r = \|\mathbf{r}\|$, so we will consider the monopole in redshift space that we denote $\xi(r)$ (and still refer to it as the correlation function as it is done in most studies)

$$\xi(r) = \frac{1}{4\pi} \int \xi(\mathbf{r}) d\Omega \quad (4)$$

In the plane parallel approximation and in the linear regime on large scales, the monopole correlation function in redshift space is linked to the correlation function in real space by a constant multiplicative factor independent of scale (Kaiser 1986).

When considering CDM models, the linear power spectrum can be computed up to an amplitude factor for given matter density ω_m , baryon density ω_b and spectral tilt n_s . In our analysis we neglect the effect of ω_b and n_s because they are well constrained by WMAP data (Komatsu et al. 2009). We fix them at the maximum likelihood values of WMAP7, $\omega_b = 2.227 \times 10^{-2}$ and $n_s = 0.966$ (we will also fix the parameter $\sigma_8 = 0.81$ for normalizing the linear power spectrum). So the only parameter of the linear power spectrum that we vary is the matter density ω_m .

A prominent feature of the linear correlation function is the BAO peak at scale ≈ 150 Mpc, which is due to sound waves traveling in the hot plasma before recombination, when photons are baryons were coupled together. Note however that the BAO peak is not the only effect of baryons in the linear correlation function, and that they also suppress the amplitude of fluctuations on small and intermediate scales.

Then we have to take into account the non-linear effects in the galaxy field. The first effect is due to the non-linear evolution of the matter density field, where recent advances in modeling have been made using Renormalized Perturbation Theory (Crocce & Scoccimarro (2006), RPT). Using RPT, it has been shown in Sánchez et al. (2008) that one can have an excellent description of the correlation function for the range of scales $60h^{-1}\text{Mpc} < r < 180h^{-1}\text{Mpc}$.

In this study we use a simple model for the non-linear evolution of the matter density field. We use the HALOFIT procedure (Smith et al. 2003), which provides corrections for scale-free power spectra using N -body simulations. Because these simulations do not include the BAO feature we also have to correct for the non-linear degradation of the acoustic peak. Eisenstein et al. (2007) found that it is well approximated by a Gaussian smoothing of the acoustic feature both in redshift and in real space.

The power spectrum with degraded peak $P_{damped,L}$ is obtained using the linear power spectrum P_L and the linear 'no wiggles' power spectrum of Eisenstein & Hu (1998), $P_{nowig,L}$

$$P_{damped,L}(k) = P_{nowig,L}(k) + e^{-a^2 k^2/2} [P_L(k) - P_{nowig,L}(k)] \quad (5)$$

To take into account the scale-free non-linear effect, we apply to the damped power spectrum the same non-linear correction as the scale-free power spectrum $P_{nowig,L}(k)$

$$P_{damped,NL}(k) = \frac{P_{NL,nowig}(k)}{P_{L,nowig}(k)} P_{damped,L}(k) \quad (6)$$

where $P_{NL,nowig}(k)$ is computed from $P_{L,nowig}(k)$ using the HALOFIT formula in Smith et al. (2003). We compute these power spectra using the iCosmo IDL library (Refregier et al. 2011).

There remains to set the value of a in formula (5) and model the scale-dependent galaxy bias with respect to the matter density field. For these purposes we use the Large Suite of Dark Matter Simulations (LasDamas, McBride et al. 2012, in prep.). These simulations are designed to model the clustering of the SDSS DR7 for galaxies in a wide luminosity range. Galaxies are artificially placed in dark matter halos using a halo occupation distribution (HOD; Berlind & Weinberg (2002)) with parameters set to match observations on the SDSS sample.

We use the gamma release of the Las Damas simulations and more precisely the Oriana simulations that are publicly available². They are composed of 40 N -body simulations, where each simulation can reproduce two times the 'North+South' SDSS footprint for a total of 80 realizations. Each N -body simulation contains 1280^3 particles of mass $45.73 \times 10^{10} h^{-1} M_\odot$ with a softening parameter of $53 h^{-1} \text{kpc}$. The cosmological parameters of the simulations are $\omega_m = 0.25$, $\Omega_\Lambda = 0.75$, $\Omega_b = 0.04$, $h = 0.7$, $\sigma_8 = 0.8$ and $n_s = 1$.

We use catalogues composed of LRG galaxies with $M_g < -21.2$ and $M_g > -23.2$ as the DR7-Full sample. As it is nearly volume-limited, the redshift range ($0.16 < z < 0.36$) is smaller than that of the DR7-Full sample. However because of a non-evolving HOD model to populate dark matter halos, the galaxy number density $n(z)$ is slowly decreasing. To address this, we compute the correlation using the random catalogue provided by the Las Damas team, which has the same decreasing trend in its density.

We compute the correlation function using the Landy-Szalay estimator of formula (11). We average the measured correlation function over the 80 realizations so that we get a very good approximation of the real correlation function. On the other hand, we compute the power spectrum as in formula (6) using the Las Damas cosmological parameters. We apply the Hankel transform to this power spectrum in order to obtain the corresponding correlation function. First we adjust the parameter a of equation (5) to reproduce the non-linear degradation in the simulations and we find that the value $a = 9.5 h^{-1} \text{Mpc}$ gives a good result. Finally we adjust the scale-dependent galaxy bias $B(r)$ on small scales by dividing the Las Damas correlation by our model. We find a scale-dependent correction of $\approx 10\%$ at $r = 5 h^{-1} \text{Mpc}$ which slowly decreases up to $r = 55 h^{-1} \text{Mpc}$.

We thus obtain the galaxy correlation function

$$\xi_{galaxy,\omega_m}(r) = B(r) \xi_{damped,NL}(r) \quad (7)$$

where $\xi_{damped,NL}(r)$ is obtained by the Hankel trans-

form of $P_{damped,NL}(k)$ of formula (6) with the choice $a = 9.5 h^{-1} \text{Mpc}$ in equation (5). We keep $B(r)$ and a fixed in our analysis, so that ξ_{galaxy,ω_m} only depends on the linear power spectra P_L and $P_{nowig,L}$ of equation (5). And as we already explained, we only vary the parameter ω_m in the linear power spectra. So the correlation function ξ_{galaxy,ω_m} only has a dependence on ω_m .

We introduce two additional parameters in the model correlation function. The first parameter α accounts for a dilation of the galaxy survey due to an incorrect choice of fiducial cosmology to convert redshifts into distances. This parameter is actually the one that is probed by the localization of the BAO peak and the standard ruler property. It was shown that a wrong choice of fiducial cosmology approximately translates into a dilation of the galaxy survey and thus of the correlation function (Eisenstein et al. 2005; Padmanabhan & White 2008) by a factor $\alpha = D_V(z_{eff})/D_{V,fid}(z_{eff})$ with $z_{eff} = 0.3$ the effective redshift of our sample, and $D_V(z)$ the 'dilation scale' at redshift z

$$D_V(z) = \left[D_M(z)^2 \frac{cz}{H(z)} \right]^{1/3} \quad (8)$$

where $H(z)$ is the Hubble parameter and $D_M(z)$ is the comoving angular diameter distance at redshift z . Our choice of a flat Λ CDM fiducial cosmology with $\Omega_m = 0.25$ gives $D_{V,fid}(z_{eff} = 0.3) = 1180 \text{ Mpc}$.

Next we introduce a constant amplitude factor b to model variations of σ_8 , linear redshift distortions and linear galaxy bias. So we obtain the final model correlation function as a function of ω_m , α and $B = b^2$

$$\xi_{\omega_m,\alpha,B}(r) = b^2 \xi_{galaxy,\omega_m}(\alpha r) \quad (9)$$

Finally we bin the model correlation function equivalently as when it is estimated by pair counting, i.e. for a bin $[r_i - dr/2, r_i + dr/2]$

$$\xi_{\omega_m,\alpha,B}(r_i) = \frac{\int_{r_i-dr/2}^{r_i+dr/2} \xi_{\omega_m h^2,\alpha,B}(r) r^2 dr}{\int_{r_i-dr/2}^{r_i+dr/2} r^2 dr} \quad (10)$$

In all this study we use a $dr = 10 h^{-1} \text{Mpc}$ binning from $20 h^{-1} \text{Mpc}$ to $200 h^{-1} \text{Mpc}$ corresponding to $n = 18$ bins.

3.2. Correlation function estimation

Most estimators of the correlation function use random unclustered catalogues (i.e. Poisson catalogues with no correlation) and compare the excess of pairs of data points separated by a distance r compared to pairs of random points. Different estimators have been proposed and compared (Pons-Bordería et al. 1999; Labatie et al. 2012a). The recommendation is to use either the Hamilton estimator (Hamilton 1993) or the Landy-Szalay estimator (Landy & Szalay 1993). They have been shown in Labatie et al. (2012a) to have lower variance than the other estimators and negligible bias for current galaxy surveys. Most studies are using the Landy-Szalay estimator, and we will also use it here. It is given by

$$\hat{\xi}(r) = 1 + \frac{N_{RR}}{N_{DD}} \frac{DD(r)}{RR(r)} - 2 \frac{N_{RR}}{N_{DR}} \frac{DR(r)}{RR(r)} \quad (11)$$

with $DD(r)$, $RR(r)$, $DR(r)$ the number of pairs at a distance in $[r \pm dr/2]$ of respectively data-data, random-

² <http://lss.phy.vanderbilt.edu/lasdamas/mocks/>

random, data-random points and N_{DD} , N_{RR} , N_{DR} the total number of corresponding pairs in the catalogues.

Formula (11) corresponds to the case where all galaxies are weighted equally in the estimator. This is optimal for volume-limited surveys but it is not optimal when the galaxy mean density depends on redshift. An approximately optimal weighting, which depends on the distance r at which we estimate the correlation function, is given in Hamilton (1993) by

$$w_i = \frac{1}{1 + \bar{n}\Phi_i J(r)} \quad (12)$$

where Φ_i is the selection function at the position of the galaxy i , \bar{n} is the expected density of the catalogue before the selection function is applied and $J(r)$ is the integral of the real correlation function

$$J(r) = \int_{V_r} \xi(\mathbf{s}) d^3\mathbf{s} = 4\pi \int_0^r \xi(s) s^2 ds \quad (13)$$

There is still a constraint not to introduce a bias, which is that the weighted density of the random catalogue and data catalogue must be proportional (i.e. there can only be a multiplicative factor of difference between the two). When introducing weights as in formula (12) the pair-counting quantities (DD , RR , DR) are modified in the Landy-Szalay estimator of equation (11). Instead of adding +1 for each pair, we simply add $w_i w_j$, with w_i and w_j the weights of each point of the pair.

When computing the correlation function of the DR7-Full sample we do not try to apply such optimal weights. We only take care of the fiber collision problem which locally changes the density of galaxies. We apply the same weights as in Kazin et al. (2010), that upweight groups of galaxies which are close enough to be affected by fiber collisions. Concerning the angular incompleteness and the varying density with redshift, they are taken into account in the random catalogue. So overall the weighted density in the data and random catalogues are proportional.

We use the same random catalogue as in Kazin et al. (2010) which is also available online³. It is composed of ≈ 1.66 million points, i.e. ≈ 16 times the number of galaxies in the data.

We plot in figure 3 the measured correlation function of the data sample, with a BAO peak a bit wider than expected. This was also found in Martínez et al. (2009) on a SDSS DR7 LRG volume-limited sample. Yet the study Kazin et al. (2010) concludes that this is not due to systematics but only to signal variance. Note also that the BAO reconstruction technique used in Padmanabhan et al. (2012) on the same sample leads to a sharpening of the BAO peak. However, without applying this technique or introducing nuisance parameters, the wide BAO peak results in a low BAO detection level and also a shift towards values $\alpha < 1$ (see section 6).

A lot of studies on the clustering of the SDSS DR7 LRG sample focused only on the position of the BAO peak. This is done either by using peak finding techniques as in Kazin et al. (2010), or by introducing nuisance parameters for the global shape of the correlation function (or

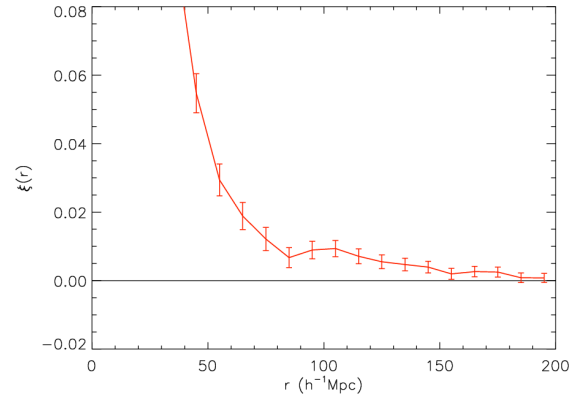


FIG. 3.— Estimated correlation function of the SDSS DR7-Full sample ξ with a flat Λ CDM fiducial cosmology with $\Omega_m = 0.25$. We give the error bars as the diagonal part $\sqrt{C_{ii}}$ of the covariance matrix obtained from 2000 lognormal simulations with parameters $\omega_m = 0.13$, $\alpha = 1$ and $b = 2.5$. The BAO peak is a bit wider than expected, which is explained by signal variance in Kazin et al. (2010).

power spectrum) which are marginalized over (e.g. spline functions in Percival et al. (2010) or inverse polynomials in Xu et al. (2012)).

In the latter case, this enables to obtain high BAO detection levels, that we do not manage to obtain here otherwise (3.6σ in Percival et al. (2010) and 3σ before reconstruction in Xu et al. (2012)). Therefore we will not study the BAO detection here. Another reason is that the presence of BAOs in large-scale structures is becoming hard to dispute after recent results from the surveys WiggleZ (3.2σ detection in Blake et al. (2011a)), 6dF (2.4σ detection in Beutler et al. (2011)) and BOSS (5σ detection in Anderson et al. (2012)). Finally let us mention that wavelet analysis also enabled to obtain high level of detection using SDSS DR7 samples (4.4σ in Arnalte-Mur et al. (2012) and 4σ in Tian et al. (2011)).

So we will focus on cosmological parameter constraints using the SDSS DR7-Full sample described in section 2. Because we use a relatively simple correlation function modeling, our study is not meant to improve cosmological parameter constraints. We only attempt to quantify the modeling error introduced by the approximation of a constant covariance C instead of a model-dependent C_θ .

4. LOGNORMAL SIMULATIONS

In this section we describe our procedure for generating lognormal simulations that will provide us with a model-dependent covariance matrix C_θ . In our lognormal simulations we use the same sky coverage and the same number density as in the SDSS DR7-Full sample.

To generate lognormal realizations we use the same method as in Labatie et al. (2012a): we generate a continuous galaxy field in a cube from an input correlation function ξ_θ , we apply the SDSS DR7-Full selection function (which incorporates the angular mask and the number density), and finally we Poisson sample the resulting continuous field.

For computational reasons we do not estimate the correlation function $\hat{\xi}$ on the full sky, but separately on the Northern Galactic Cap, $\hat{\xi}_{NGC}$ and Southern Galactic Cap, $\hat{\xi}_{SGC}$, which can be considered as independent. Also for computational reasons we use random catalogues with the same density as the SDSS DR7-Full sample.

³ <http://cosmo.nyu.edu/~eak306/SDSS-LRG.html>

From these measurements we obtain the model-dependent covariance matrices $C_{NGC,\theta}$, and $C_{SGC,\theta}$ by computing the empirical covariance matrices (as in equations (1) and (2)). For each simulation, corresponding to a parameter θ , we obtain the full correlation function $\hat{\xi}$ by the same optimal linear combination as in White et al. (2011) (see appendix A)

$$\hat{\xi} = C_\theta \left[C_{NGC,\theta}^{-1} \hat{\xi}_{NGC} + C_{SGC,\theta}^{-1} \hat{\xi}_{SGC} \right] \quad (14)$$

$$C_\theta = \left(C_{NGC,\theta}^{-1} + C_{SGC,\theta}^{-1} \right)^{-1} \quad (15)$$

with C_θ the resulting covariance matrix of the full correlation $\hat{\xi}$.

As we stated in section 3.1 we only take into account 3 main parameters in the correlation function, i.e. $\theta = (\omega_m, \alpha, B)$.

The parameter ω_m changes the whole shape of the correlation function, so we have no choice but to generate different sets of lognormal simulations for different values of ω_m . We choose to use 5 values $\omega_m = 0.08, 0.105, 0.13, 0.155, 0.18$ and simply interpolate linearly the covariance matrix for intermediate values (more precisely, each coefficient of the covariance matrix is linearly interpolated).

The parameter α , on the other hand, only creates a dilation of the galaxy survey and thus of the apparent correlation function. This is only a geometrical effect due to a wrong fiducial cosmology. It is thus possible to take it into account using a single set of simulations.

First we must take into account that if the survey extends from a minimum distance r_{min} to a maximum distance r_{max} in fiducial coordinates, it extends from αr_{min} to αr_{max} in comoving coordinates. So for a simulation parameter α , one must consider cuts at these distances αr_{min} and αr_{max} and then artificially dilate the survey by a factor α to mimic the effect of a wrong fiducial cosmology.

So instead of producing simulations that extend from r_{min} to r_{max} , we produce simulations that extend from $\alpha_{min} r_{min}$ to $\alpha_{max} r_{max}$, where α_{min} and α_{max} are the minimum and maximum values of α considered. In this way we are always able to consider cuts at distances αr_{min} and αr_{max} . In this study we choose $\alpha_{min} = 0.8$ and $\alpha_{max} = 1.2$. Given the value $D_{V,fid}(0.3) = 1180$ Mpc for our fiducial cosmology, we get a probed range $D_V(0.3) \in [944 \text{ Mpc}, 1416 \text{ Mpc}]$.

There is another complication because the apparent density must be in agreement with the one observed in the data catalogue. So in addition to the cuts between αr_{min} and αr_{max} , we introduce a varying selection function that depends on α , so that the observed density after the dilation by α agrees with the one of the data catalogue.

We developed an optimized procedure for computing the correlation function in this context. First, because the correlation function is estimated by pair-counting, the estimation can be done in comoving coordinates (i.e. before the dilation) and the dilation is only applied after the pair-counting by dilating bin ranges. The density in comoving space is given by

$$n_\alpha(r) = \frac{1}{\alpha^3} n\left(\frac{r}{\alpha}\right) \quad (16)$$

with $n(r)$ the observed density in the data catalogue and the factor $1/\alpha^3$ accounting for the change of density because of the dilation.

So the original lognormal simulations are generated with a density $n_{max}(r) = \max_\alpha n_\alpha(r)$. Let us define the selection function $\Phi_\alpha(r) = n_\alpha(r)/n_{max}(r)$. We apply this selection function for every value of α in the following way: for each galaxy at distance r in the original simulation, we generate a random uniform variable $u \in [0, 1]$. Then the galaxy belongs to the simulation with value α if $u < \Phi_\alpha(r)$.

For each galaxy \mathbf{x}_i we end up with a sequence of intervals $[\alpha_i, \alpha'_i]$ for which the galaxy belongs to the simulations. To optimize the computation of the correlation function we create a new galaxy at the same position for every distinct interval $[\alpha_i, \alpha_{i+1}]$.

Let us consider only the pair counting term DD , with the same argument that could be applied for DR and RR . For every r we consider an array $(DD_{\alpha_i, raw}(r))_{i=1, \dots, n}$ corresponding to the grid $\alpha = (\alpha_1, \dots, \alpha_n)$. This counts the number of pairs to add from $DD_{\alpha_i}(r)$ to obtain $DD_{\alpha_{i+1}}(r)$.

For every pair $(\mathbf{x}_k, \mathbf{x}_l)$ with α ranges respectively equal to $[\alpha_k, \alpha_{k'}]$ and $[\alpha_l, \alpha_{l'}]$, the pair belongs to the simulations for the range $[\max(\alpha_k, \alpha_l), \min(\alpha_{k'}, \alpha_{l'})] = [\alpha_{\max(k,l)}, \alpha_{\min(k',l')}]$. So we add +1 to $DD_{\alpha, raw}(r)$ for $\alpha = \alpha_{\max(k,l)}$ and add -1 for $\alpha = \alpha_{\min(k',l') + 1}$. In the end we obtain the α dependent $DD_\alpha(r)$ as

$$DD_{\alpha_i}(r) = \sum_{j=0}^i DD_{\alpha_j, raw}(r) \quad (17)$$

Finally we only have to perform the dilation on $DD_{\alpha_i}(r)$ that was computed in comoving coordinates

$$DD_\alpha^{final}(r) = DD_\alpha(\alpha r) \quad (18)$$

This whole procedure enables to obtain DD , DR and RR for every r and every α with a time increased only by a factor ≈ 4 instead of being proportional to the number of α values.

Finally let us turn to the third parameter $B = b^2$, which changes the real galaxy distribution in comoving space, just like ω_m . But because it is simply a constant multiplicative factor B in the correlation function, it should give approximately a factor B^2 in the covariance matrix of $\hat{\xi}$. We recall that there are two different sources of noise in the estimator $\hat{\xi}$

- Cosmic variance due to the finite extent of the catalogue
- Shot noise due to the finite number of galaxies to map an underlying continuous field

The approximation of a covariance matrix scaling as B^2 is valid when we can neglect the shot noise contribution compared to the cosmic variance contribution. So obviously it is better verified for large values of b . However we verify in section 5.3 that it is a good approximation around reasonable values of b , with the approximation $B^2 C$ being much closer to the real covariance matrix than the approximation of a constant C . So this parameter will actually be treated without any need for more simulations.

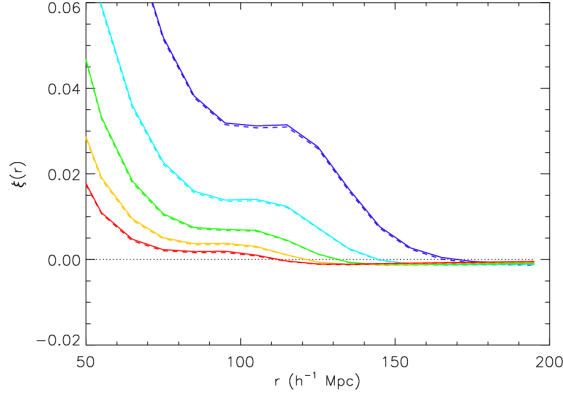


FIG. 4.— Mean estimators $\bar{\xi}_{\omega_m}$ in dashed lines compared to the real correlation function ξ_{ω_m} in solid lines for $\alpha = 1$ and for $\omega_m = 0.08$ (purple), 0.105 (light blue), 0.13 (green), 0.155 (yellow), 0.18 (red).

Our main set of simulations will be performed with $b = 2.5$ (note that this value is with respect to the real space correlation, i.e. without the boost factor of Kaiser (1986)). For each value of (ω_m, α) we will use $N = 2000$ lognormal simulations to estimate the covariance matrix $C_{\omega_m, \alpha}$.

5. RESULTS ON THE STATISTICAL MODELING OF $\hat{\xi}$

5.1. Absence of bias in $\hat{\xi}$

We first test whether there is a bias affecting the estimators of the correlation function in our lognormal simulations. This is important for cosmological parameter constraints because the expected value of $\hat{\xi}$ is assumed to be from a given model ξ_θ (see section 6)

$$\exists \theta \in \Theta \text{ s.t. } \hat{\xi} \sim \mathcal{N}(\xi_\theta, C_\theta) \quad (19)$$

To verify that the bias is negligible we compute the mean of the measured correlation function for $\alpha = 1$ and for the different values $\omega_m = 0.08, 0.105, 0.13, 0.155, 0.18$, using $N = 2000$ lognormal simulations in each case

$$\bar{\xi}_{\omega_m} = \frac{1}{N} \sum_{k=1}^N \hat{\xi}_{k, \omega_m} \quad (20)$$

We plot in figure 4 the resulting mean estimators $\bar{\xi}_{\omega_m}$ compared to the real correlation function ξ_{ω_m} , which is given as the lognormal simulations input. Figure 4 shows a very good agreement, i.e. that the estimators are nearly unbiased.

5.2. Verification of the Gaussianity of $\hat{\xi}$

Now we want to verify the Gaussianity of the measured correlation function $\hat{\xi}$, i.e. again to verify that the following hypothesis is realistic

$$\exists \theta \in \Theta \text{ s.t. } \hat{\xi} \sim \mathcal{N}(\xi_\theta, C_\theta)$$

For this we use the correlation function estimates $\hat{\xi}$ on the $N = 80$ Las Damas realizations presented in section 3.1. Indeed they are more realistic than our lognormal realizations. For example the broadening of the BAO feature appears through non-linear evolution in the Las Damas simulations, whereas it is simply 'injected' through the input correlation function in our lognormal realizations.

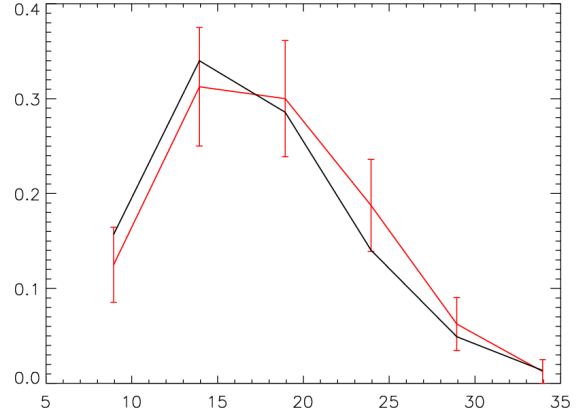


FIG. 5.— Estimated pdf of χ^2 (red) using the histogram on the 80 Las Damas realizations and pdf of a χ^2_{18} distribution (black). Error bars give the Poisson uncertainty in the estimate due to finite number of realizations.

First we compute the empirical mean and empirical covariance matrix of the LasDamas realizations

$$\bar{\xi} = \frac{1}{N} \sum_{k=1}^N \hat{\xi}_k \quad (21)$$

$$C_{ij} = \frac{1}{N-1} \sum_{k=1}^N [\hat{\xi}_k(r_i) - \bar{\xi}(r_i)] [\hat{\xi}_k(r_j) - \bar{\xi}(r_j)] \quad (22)$$

We then compute the χ^2 statistic for each realization $\hat{\xi}_k$, which should approximately follow a χ^2_n law with $n = 18$ if the measurement $\hat{\xi}$ is Gaussian

$$\chi^2 = \langle \hat{\xi} - \bar{\xi}, C^{-1}(\hat{\xi} - \bar{\xi}) \rangle \quad (23)$$

$$= \sum_{1 \leq i, j \leq n} [\hat{\xi}(r_i) - \bar{\xi}(r_i)] C_{i,j}^{-1} [\hat{\xi}(r_j) - \bar{\xi}(r_j)] \quad (24)$$

We show on figure 5 the histogram of χ^2 on the 80 Las Damas realizations compared to the probability density function (pdf) of a χ^2_n variable with $n = 18$. We can see the very good agreement between the two distributions.

5.3. Dependence of C_θ on ω_m , α and B

Here we describe the dependence of C_θ (obtained from our full set of lognormal simulations) with respect to ω_m , α and B .

First we check that the dependence of C_θ on $B = b^2$ can actually be approximated as $C_{\omega_m, \alpha, B} \propto B^2 C_{\omega_m, \alpha}$. For this we compare the covariance matrix $C_1 = C_{\omega_m, \alpha, B_1}$ to the covariance matrix $C_2 = C_{\omega_m, \alpha, B_2}$ obtained in each case from $N = 2000$ lognormal simulations, respectively with $\omega_m = 0.13$, $\alpha = 1$, $B_1 = 2.5^2$ and $\omega_m = 0.13$, $\alpha = 1$, $B_2 = 3.0^2$.

We compute the L2 distance between $(B_2/B_1)^2 C_1$ and C_2 , and compare it to the L2 distance between C_1 and C_2

$$\frac{\|(B_2/B_1)^2 C_1 - C_2\|_2}{\|C_1 - C_2\|_2} = 0.22 \quad (25)$$

So we obtain that the approximation $C_{\omega_m, \alpha, B} \propto B^2 C_{\omega_m, \alpha}$ is 5 times better than the approximation of a constant covariance matrix, which justifies our approximation.

Next we outline the significant dependence of C_θ with respect to the two other parameters ω_m and α . We start by analyzing the dependence of C_θ with respect to ω_m in the case $\alpha = 1$ and $B = 2.5^2$. We show on figure 6 and 7 the variations of C_θ for $\omega_m = 0.08, 0.105, 0.13, 0.155, 0.18$. For clarity reasons we distinguish between the correlation matrix ρ_θ (i.e. the covariance matrix normalized by the diagonal elements) of formula (26) and the diagonal part $\sigma_\theta = (\sqrt{C_{\theta,ii}})$, which fully describe the covariance matrix together.

$$\rho_{\theta,ij} = \frac{C_{\theta,ij}}{\sqrt{C_{\theta,ii}C_{\theta,jj}}} = \frac{1}{\sigma_{\theta,i}\sigma_{\theta,j}}C_{\theta,ij} \quad (26)$$

We recall that the correlation function has $n = 18$ bins of size $dr = 10h^{-1}\text{Mpc}$ from $20h^{-1}\text{Mpc}$ to $200h^{-1}\text{Mpc}$. We find a strong dependence of the whole covariance matrix with respect to ω_m , i.e. both the diagonal part σ_θ and the correlation matrix ρ_θ have a strong dependence on ω_m .

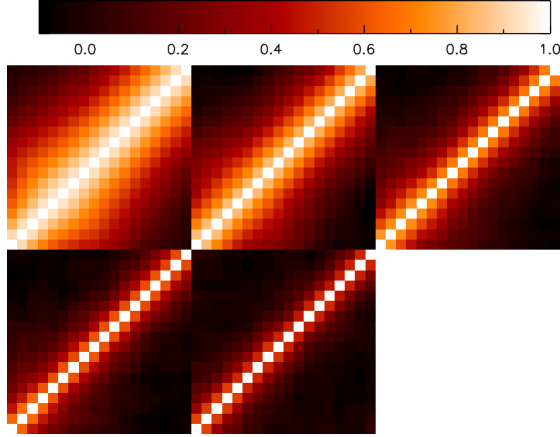


FIG. 6.— Dependence of ρ_θ with respect to ω_m , in the case $\alpha = 1$ and $B = 2.5^2$. We plot ρ_θ for $\omega_m = 0.08$ (top left), 0.105 (top middle), 0.13 (top right), 0.155 (bottom left), 0.18 (bottom middle). The correlation between bins strongly increases for smaller values of ω_m . We have plotted the $n = 18$ bins of size $dr = 10h^{-1}\text{Mpc}$ from $20h^{-1}\text{Mpc}$ to $200h^{-1}\text{Mpc}$.

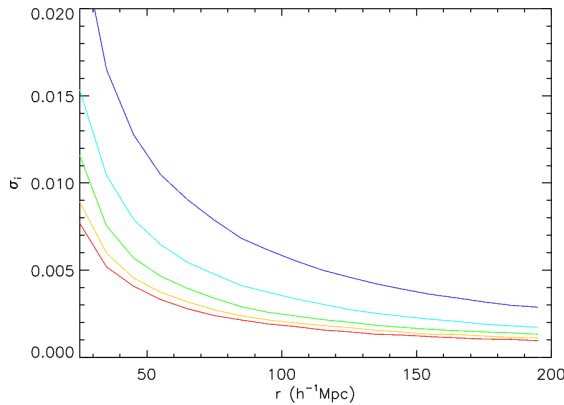


FIG. 7.— Dependence of $\sigma_\theta = (\sqrt{C_{\theta,ii}})$ with respect to ω_m , in the case $\alpha = 1$ and $B = 2.5^2$. We plot σ_θ for $\omega_m = 0.08$ (purple), 0.105 (light blue), 0.13 (green), 0.155 (yellow), 0.18 (red). The diagonal variance strongly increases for smaller values of ω_m .

Finally we analyze the dependence of C_θ with re-

spect to α in the case $\omega_m = 0.13$ and $B = 2.5^2$. We show on figure 8 and 9 the variations of C_θ for $\alpha = 0.8, 0.9, 1.0, 1.1, 1.2$, again plotting separately the correlation matrix ρ_θ and the diagonal part σ_θ . We also find a dependence of both ρ_θ and σ_θ with respect to α but this dependence is not as strong as for ω_m . Note that this conclusion is dependent on the ranges of parameter values, but here we considered pretty standard ranges.

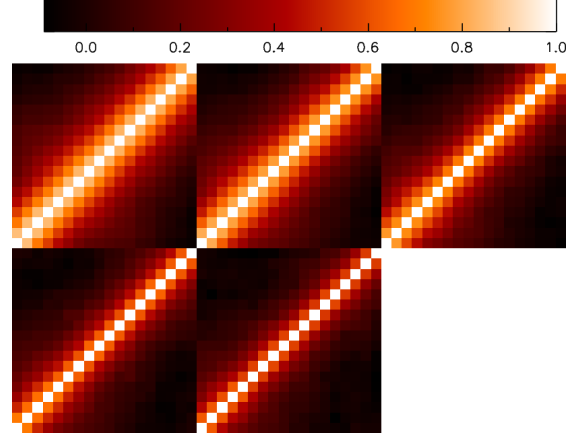


FIG. 8.— Dependence of ρ_θ with respect to α , in the case $\omega_m = 0.13$ and $B = 2.5^2$. We plot ρ_θ for $\alpha = 0.8$ (top left), 0.9 (top middle), 1.0 (top right), 1.1 (bottom left), 1.2 (bottom middle). The correlation between bins increases for smaller values of α . We have plotted the $n = 18$ bins of size $dr = 10h^{-1}\text{Mpc}$ from $20h^{-1}\text{Mpc}$ to $200h^{-1}\text{Mpc}$.

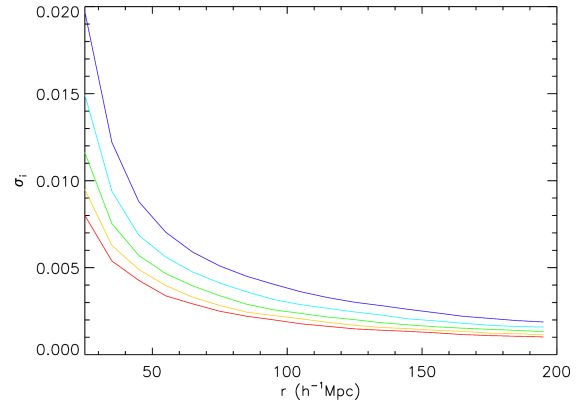


FIG. 9.— Dependence of σ_θ with respect to α , in the case $\omega_m = 0.13$ and $B = 2.5^2$. We plot σ_θ for $\alpha = 0.8$ (purple), 0.9 (light blue), 1.0 (green), 1.1 (yellow), 1.2 (red). The variance increases for smaller values of α .

6. EFFECT OF C_θ FOR COSMOLOGICAL PARAMETER CONSTRAINTS

To obtain cosmological parameter constraints from BAOs one usually perform a likelihood analysis using the whole correlation function (Eisenstein et al. 2005; Sánchez et al. 2009; Beutler et al. 2011; Blake et al. 2011a,b) or power spectrum (Cole et al. 2005; Tegmark et al. 2006; Padmanabhan et al. 2007; Reid et al. 2010; Ho et al. 2012), though some studies effectively restrict the analysis to the position of the BAO peak (Percival et al. 2007, 2010; Kazin et al. 2010; Mehta et al. 2012).

One supposes that the following hypothesis is true and wants to constrain the parameter θ

$$\exists \theta \in \Theta \text{ s.t. } \hat{\xi} \sim \mathcal{N}(\xi_\theta, C_\theta)$$

To obtain posterior information on θ one needs a Bayesian point of view by assuming a prior $p(\theta)$. Then the posterior of θ knowing the measurement $\hat{\xi}$ is given by the Bayes' theorem

$$p(\theta | \hat{\xi}) \propto p(\theta) p(\hat{\xi} | \theta) = p(\theta) \mathcal{L}_\theta(\hat{\xi}) \quad (27)$$

The combination of the measurement $\hat{\xi}$ with other independent experiments can be done inside the prior. For example with CMB data the posterior is given by

$$\begin{aligned} p(\theta | \text{CMB}, \hat{\xi}) &\propto p(\theta, \text{CMB}, \hat{\xi}) \\ &\propto p(\theta, \text{CMB}) p(\hat{\xi} | \theta, \text{CMB}) \\ &\propto p(\theta | \text{CMB}) \mathcal{L}_\theta(\hat{\xi}) \end{aligned} \quad (28)$$

where we used the independence of $\hat{\xi}$ and CMB measurement. Adding the CMB measurement is thus equivalent to using a prior $p(\theta) = p(\theta | \text{CMB})$.

To constrain θ only from the measurement $\hat{\xi}$ the question of choosing a prior $p(\theta)$ can be difficult. In this study we take a constant prior $p(\theta)$, but note that this choice is arbitrary. So the posterior is equivalent to the likelihood

$$\mathcal{L}_\theta(\hat{\xi}) \propto |C_\theta|^{-1/2} e^{-\frac{1}{2}(\hat{\xi} - \xi_\theta, C_\theta^{-1}(\hat{\xi} - \xi_\theta))} \quad (29)$$

In all the following we compare the posterior obtained using our model-dependent C_θ to the posterior obtained with constant covariance matrix $C = C_{\theta_0}$ for the particular value $\theta_0 = (\omega_m, \alpha, B) = (0.13, 1.0, 2.5^2)$. We only plot the 2D posteriors $p(\omega_m, D_V(0.3) | \hat{\xi})$ (we recall the simple relation $\alpha = D_V(0.3)/D_{V, fid}(0.3)$), i.e. after marginalizing over $B = b^2$.

$$p(\omega_m, D_V(0.3) | \hat{\xi}) = \int_B p(\omega_m, D_V(0.3), B | \hat{\xi}) dB \quad (30)$$

where we will consider the following grid: $B \in [4.0, 9.0]$ with grid step $dB = 0.01$, $\omega_m \in [0.08, 0.18]$ with grid step 0.00025 and $\alpha \in [0.8, 1.2]$ with grid step 0.001. This grid in α corresponds to a grid $D_V(0.3) \in [944 \text{ Mpc}, 1416 \text{ Mpc}]$ with grid step 1.18 Mpc.

6.1. Effect of C_θ on the SDSS DR7-Full $\hat{\xi}$

Here we work with the SDSS DR7-Full estimated correlation function $\hat{\xi}$ of figure 3.

We plot in figures 10 and 11 the posterior $p(\omega_m, D_V(0.3) | \hat{\xi})$, respectively for a constant covariance matrix $C = C_{\theta_0}$ and for a model-dependent covariance matrix C_θ .

First we can notice that the posterior $p(\omega_m, D_V(0.3) | \hat{\xi})$ is less regular and more 'noisy' in the case of model-dependent C_θ . This can be easily explained by the noise in the estimation of C_θ .

We also notice that the 2-dimensional posterior cannot be so well approximated by a 2-dimensional Gaussian (characterized notably by elliptical contours), especially in the case of constant C . We attribute this to the behavior of the model correlation function ξ_θ for high ω_m and low α (bottom right of figure 10).

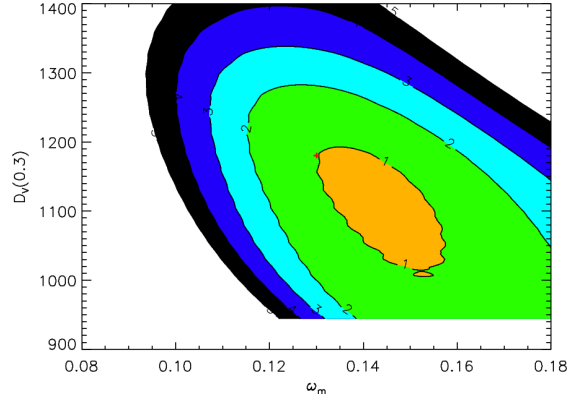


FIG. 10.— Posterior $p(\omega_m, D_V(0.3) | \hat{\xi})$ in the case of constant covariance matrix $C = C_{\theta_0}$, with $\theta_0 = (\omega_m, \alpha, B) = (0.13, 1.0, 2.5^2)$ (position of the red cross on the figure), for the SDSS DR7-Full measurement $\hat{\xi}$. We plot the 1σ to 5σ confidence regions with the approximation that p is a 2-dimensional Gaussian. They correspond respectively to $-2\ln(p) = -2\ln(p_{max}) + 2.29, 6.16, 11.81, 19.32, 28.74$ (see section 'Confidence Limits on Estimated Model Parameters' in Press et al. (2007)). We obtain the 1-dimensional constraints $\omega_m = 0.145 \pm 0.016$ (10.8% precision) and $D_V(0.3) = 1104 \pm 105 \text{ Mpc}$ (9.5% precision).

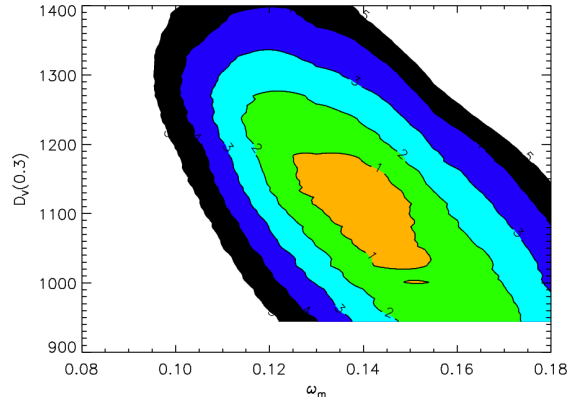


FIG. 11.— Posterior $p(\omega_m, D_V(0.3) | \hat{\xi})$ in the case of model-dependent covariance matrix C_θ for the SDSS DR7-Full measurement $\hat{\xi}$. We obtain the 1-dimensional constraints $\omega_m = 0.140 \pm 0.011$ (7.9% precision) and $D_V(0.3) = 1114 \pm 74 \text{ Mpc}$ (6.7% precision). There is a small shift in the position of the posterior's maximum and the confidence regions get smaller when considering a model-dependent C_θ .

From the 2-dimensional posteriors we compute 1-dimensional posteriors on ω_m and $D_V(0.3)$, by marginalizing over the other parameter. Then we compute 1-dimensional constraints, that we express as a symmetric 68% confidence interval (1σ interval) around the posterior's maximum.

In the case of constant covariance matrix C , we obtain the constraints $\omega_m = 0.145 \pm 0.016$ (10.8% precision) and $D_V(0.3) = 1104 \pm 105 \text{ Mpc}$ (9.5% precision). Whereas in the case of model-dependent covariance matrix C_θ , we obtain the constraints $\omega_m = 0.140 \pm 0.011$ (7.9% precision) and $D_V(0.3) = 1114 \pm 74 \text{ Mpc}$ (6.7% precision). In terms of α , this gives respectively the constraints $\alpha = 0.935 \pm 0.089$ for constant C and $\alpha = 0.944 \pm 0.063$ for model-dependent C_θ .

As can be seen when comparing figures 10 and 11 the modeling error due to the approximation of constant C is relatively small. Compared to the size of the 1σ intervals, the maximum likelihood positions are shifted by

respectively 31% for ω_m and 10% for α . The 1σ intervals also get reduced by respectively 31% for ω_m and 29% for α . However we will see in section 6.2 that the reduction of the 1σ region is not systematic.

6.2. Quantifying the effect of C_θ on SDSS DR7-Full simulations

The approximation of C_θ as a constant C results in a modeling error, which potentially depends on the particular realization $\hat{\xi}$. So we want to quantify the general effect of this approximation on cosmological parameter constraints using a lot of realizations $\hat{\xi}$

$$\hat{\xi} \sim \mathcal{N}(\xi_{\theta_0}, C_{\theta_0}) \quad (31)$$

with the choice $\theta_0 = (\omega_m, \alpha, B) = (0.13, 1.0, 2.5^2)$. For each realization $\hat{\xi}$ we compute the 2-dimensional posterior $p(\omega_m, \alpha | \hat{\xi})$ in the case of constant C and model-dependent C_θ .

We look at two particular modeling errors

- Error on the position of the 1-dimensional posterior's maxima ω_m^{\max} and α^{\max}
- Error on the size of the 1σ intervals σ_{ω_m} and σ_α

We adopt the following notations

$$\delta\omega_m^{\max} = (\omega_m^{\max, C} - \omega_m^{\max, C_\theta}) / \sigma_{\omega_m}^C \quad (32)$$

$$\delta\sigma_{\omega_m} = (\sigma_{\omega_m}^C - \sigma_{\omega_m}^{C_\theta}) / \sigma_{\omega_m}^C \quad (33)$$

$$\delta\alpha^{\max} = (\alpha^{\max, C} - \alpha^{\max, C_\theta}) / \sigma_\alpha^C \quad (34)$$

$$\delta\sigma_\alpha = (\sigma_\alpha^C - \sigma_\alpha^{C_\theta}) / \sigma_\alpha^C \quad (35)$$

We generate 2000 realizations following the model of formula (31) and look at the different quantities $\delta\omega_m^{\max}$, $\delta\sigma_{\omega_m}$, $\delta\alpha^{\max}$ and $\delta\sigma_\alpha$, which characterize the modeling error due to incorrect covariance matrix for each realization $\hat{\xi}$. Each quantity is divided by the 1σ interval size (the statistical uncertainty) in equations (32), (33), (34) and (35) in order to compare the modeling error to the statistical uncertainty.

We compute the mean values $\langle\delta\omega_m^{\max}\rangle$, $\langle\delta\sigma_{\omega_m}\rangle$, $\langle\delta\alpha^{\max}\rangle$ and $\langle\delta\sigma_\alpha\rangle$ to investigate a systematic shift in the posterior's maxima or a systematic reduction of the 1σ intervals. However we found that these mean values are negligible compared to the 1σ interval sizes ($\approx 2\%$).

Next we compute the mean absolute values $\langle|\delta\omega_m^{\max}|\rangle$, $\langle|\delta\sigma_{\omega_m}|\rangle$, $\langle|\delta\alpha^{\max}|\rangle$ and $\langle|\delta\sigma_\alpha|\rangle$. $\langle|\delta\omega_m^{\max}|\rangle$ and $\langle|\delta\alpha^{\max}|\rangle$ give the mean modeling error on the position of the posterior's maxima compared to the 1σ interval sizes. On the other hand, $\langle|\delta\sigma_{\omega_m}|\rangle$ and $\langle|\delta\sigma_\alpha|\rangle$ give the mean modeling error on the 1σ interval sizes. These absolute values actually correspond to what is normally referred as the modeling error (indeed for a given realization $\hat{\xi}$, we do not really care about the sign of the error but only on its amplitude). We show our results in table 2.

As shown in table 2, there is a mean modeling error of 21% to 28% for the position of the posterior's maxima and 7.5% to 10% for the size of the 1σ intervals. So the position of the posterior's maxima is much more affected by the modeling error than the 1σ intervals. However the error stays quite small compared to the 1σ intervals.

TABLE 2

$\langle \delta\omega_m^{\max} \rangle$	21%
$\langle \delta\sigma_{\omega_m} \rangle$	7.5%
$\langle \delta\alpha^{\max} \rangle$	28%
$\langle \delta\sigma_\alpha \rangle$	10%

NOTES.—Importance of the modeling error compared to the 1σ interval size, both for the position of the posterior's maxima and for the size of 1σ intervals. We find a mean modeling error which is quite small compared to the 1σ interval sizes.

From table 2, we see that the error on the extremities of the 1σ intervals is likely to stay below $21\% + 7.5\% = 28.5\%$ for ω_m and $28\% + 10\% = 38\%$ for α . So a possible way to handle the modeling error (though it cannot be handled for sure, because it depends on the particular realization $\hat{\xi}$) is to multiply the size of 1σ intervals obtained with a constant covariance matrix C by a factor ≈ 1.3 for ω_m and ≈ 1.4 for α . In this way, the new 1σ intervals will very likely cover most of the real 1σ intervals (i.e. the ones obtained with a model-dependent C_θ).

Let us illustrate more clearly how the modeling error can vary depending on the realization $\hat{\xi}$. On figure 12 we show for each quantity $\delta\omega_m^{\max}$, $\delta\sigma_{\omega_m}$, $\delta\alpha^{\max}$ and $\delta\sigma_\alpha$ the estimated probability density function (pdf) from their histogram on 2000 realizations $\hat{\xi}$. We clearly see that the small modeling error varies depending on the realization $\hat{\xi}$.

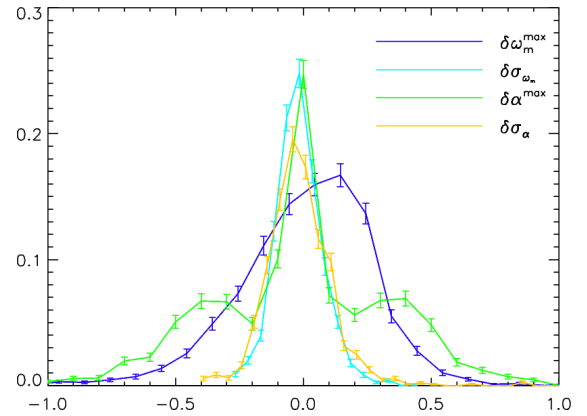


FIG. 12.— Estimated pdf of $\delta\omega_m^{\max}$, $\delta\sigma_{\omega_m}$, $\delta\alpha^{\max}$ and $\delta\sigma_\alpha$ using their histogram on 2000 realizations. Error bars give the Poisson uncertainty in the estimate due to finite number of realizations.

Finally we perform a visual inspection of the 2-dimensional posteriors $p(\omega_m, D_V(0.3) | \hat{\xi})$ in both cases of constant covariance C and model-dependent C_θ . As in section 6.1 we find deviations of the 2-dimensional posteriors compared to 2-dimensional Gaussians for most realizations $\hat{\xi}$. These deviations are located at high ω_m and low α and they happen both in the case of constant C and model-dependent C_θ . So they are simply due to the behavior of the model correlation function ξ_θ in this region.

6.3. Quantifying the effect of C_θ for next-generation surveys

Finally we try to quantify this modeling error for next-generation surveys. Our procedure is simply to divide

the covariance matrices C and C_θ by a constant factor c with $c = 2$ and $c = 4$, and repeat the analysis of section 6.2. To give an idea of what this represents in terms of survey size, we can approximate doubling the survey size as a factor 1/2 in the covariance matrix

$$C \left[\frac{1}{2} \hat{\xi}_1 + \frac{1}{2} \hat{\xi}_2 \right] \approx \frac{1}{4} C \left[\hat{\xi}_1 \right] + \frac{1}{4} C \left[\hat{\xi}_2 \right] \quad (36)$$

$$\approx \frac{1}{2} C \left[\hat{\xi}_1 \right] \quad (37)$$

because the estimated correlation function $\hat{\xi}_{12}$ of survey '1+2' is approximately the same as the mean of $\hat{\xi}_1$ and $\hat{\xi}_2$ for large enough surveys. So a factor 1/2 in the covariance matrix is approximately equivalent to doubling the survey size, and similarly a factor 1/4 in the covariance matrix is approximately equivalent to quadrupling the survey size.

Now we generate realizations from the model

$$\hat{\xi} \sim \mathcal{N} \left(\xi_{\theta_0}, \frac{1}{c} C_{\theta_0} \right) \quad (38)$$

The approximate likelihood (with constant covariance matrix) and real likelihood (with model-dependent covariance matrix) are now given by

$$\mathcal{L}_\theta^C(\hat{\xi}) \propto e^{-\frac{c}{2} \langle \hat{\xi} - \xi_\theta, C^{-1}(\hat{\xi} - \xi_\theta) \rangle} \quad (39)$$

$$\mathcal{L}_\theta^{C_\theta}(\hat{\xi}) \propto |C_\theta|^{-1/2} e^{-\frac{c}{2} \langle \hat{\xi} - \xi_\theta, C_\theta^{-1}(\hat{\xi} - \xi_\theta) \rangle} \quad (40)$$

We repeat the analysis of table 2 with 2000 realizations of formula (38) for each case $c = 2$ and $c = 4$. We show the results in table 3.

TABLE 3

	$c = 2$	$c = 4$
$\langle \delta\omega_m^{\max} \rangle$	16%	13%
$\langle \delta\sigma_{\omega_m} \rangle$	6.3%	5.1%
$\langle \delta\alpha^{\max} \rangle$	23%	20%
$\langle \delta\sigma_\alpha \rangle$	8.5%	6.9%

NOTES.—Importance of the modeling error compared to the 1σ intervals size, both for the position of the posterior's maxima and for the size of 1σ intervals when dividing C and C_θ by factors $c = 2$ and $c = 4$. Again we find a mean modeling error which is quite small compared to the 1σ interval sizes. The error is smaller here than for the SDSS DR7-Full simulations, and it decreases with the survey size.

From table 3 we find again that there is mean modeling error which is quite small compared to the 1σ interval sizes. The modeling error mainly affects the position of the posterior's maxima. It is smaller here than for the SDSS DR7-Full simulations, and it decreases with the survey size.

Our conclusion is that the approximation of C_θ as a constant C only has a small impact on cosmological parameter constraints. As surveys get larger the modeling error decreases. Again an approximate way to handle this modeling error is to multiply the size of 1σ intervals by a factor ≈ 1.3 . We emphasize that our study is

not comprehensive and that we only took into account 3 parameters: $\theta = (\omega_m, \alpha, B)$.

This conclusion is a bit surprising since we found a strong dependence of C_θ on θ . However it is easy to see that there is a competing effect at work in the likelihood. Let us remind the expression of the likelihood for a model-dependent covariance matrix C_θ

$$\mathcal{L}_\theta(\hat{\xi}) \propto |C_\theta|^{-1/2} e^{-\frac{1}{2} \langle \hat{\xi} - \xi_\theta, C_\theta^{-1}(\hat{\xi} - \xi_\theta) \rangle} \quad (41)$$

For example if we multiply the covariance matrix by a constant factor c the terms $|C_\theta|^{-1/2}$ and $e^{-\frac{1}{2} \langle \hat{\xi} - \xi_\theta, C_\theta^{-1}(\hat{\xi} - \xi_\theta) \rangle}$ will have competing effects, decreasing the overall effect on the likelihood. And we indeed verified that the term $|C_\theta|^{-1/2}$ has an important contribution in practice (i.e. if it is omitted, one obtains much greater changes of the likelihood contours).

Finally we also perform a visual inspection of the 2-dimensional posteriors $p(\omega_m, D_V(0.3) | \hat{\xi})$ in both cases of constant covariance C and model-dependent C_θ for $c = 2$ and $c = 4$. Because the maximum likelihood is much closer to the real parameter θ_0 of formula (38) than in section 6.2 (because variations of $\hat{\xi}$ are smaller), the region causing deviations to a 2-dimensional Gaussian is nearly always outside the 2 to 3σ confidence region. So we find that realizations $\hat{\xi}$ of formula (38) have 2-dimensional posteriors that can be very well approximated by 2-dimensional Gaussians.

7. CONCLUSIONS

In this paper we have studied the influence of considering a realistic model-dependent covariance matrix C_θ instead of a constant covariance matrix C of the estimated correlation function $\hat{\xi}$ for cosmological parameter constraints. The main difficulty comes from the very long computation time required to estimate such a model-dependent covariance matrix C_θ .

We have presented a new method to obtain a realistic model-dependent C_θ in a reasonable time, for a 3-dimensional parameter $\theta = (\omega_m, \alpha, b^2)$ using lognormal simulations. Compared to a constant covariance matrix, the computing time is multiplied by a factor roughly 20. We plan to release (as part of a general toolbox on the correlation function analysis of galaxy clustering) the different programs to estimate a model-dependent C_θ for different survey masks, selection functions and ranges of cosmological parameters.

Our first results concern the statistical modeling of the measured correlation function $\hat{\xi}$

$$\exists \theta \in \Theta \text{ s.t. } \hat{\xi} \sim \mathcal{N}(\xi_\theta, C_\theta) \quad (42)$$

We verified the absence of bias in our lognormal simulations, i.e. that the expected value of measured correlation function $\mathbb{E}(\hat{\xi})$ is indeed equal to the input model in our simulations ξ_θ . Next we verified the Gaussianity of the measurement $\hat{\xi}$ using 80 Las Damas realizations, which are more realistic than our lognormal simulations. We estimated the probability density function of a χ^2 statistic on these 80 realizations, and found that it is compatible with the expected result for Gaussian realizations.

We also studied the dependence of C_θ with respect to ω_m , α and $B = b^2$. We found that the effect of the amplitude parameter b^2 can be well approximated as a constant factor b^4 in the covariance matrix (for b high enough, i.e. > 2). For the two other parameters ω_m and α , we found that their variations affect the whole shape of the covariance matrix. However ω_m has a bigger effect than α for usual ranges of parameter values.

Next we studied the implications of a model-dependent C_θ for cosmological parameter constraints. More precisely, we always compared the results obtained with C_θ to the results obtained with a constant $C = C_{\theta_0}$ for the particular value $\theta_0 = (\omega_m, \alpha, b^2) = (0.13, 1.0, 2.5^2)$.

For the SDSS DR7-Full sample, we obtained $\omega_m = 0.145 \pm 0.016$ (10.8% precision) and $D_V(0.3) = 1104 \pm 105$ Mpc (9.5% precision) for a constant C , whereas we obtained $\omega_m = 0.140 \pm 0.011$ (7.9% precision) and $D_V(0.3) = 1114 \pm 74$ Mpc (6.7% precision) for a model-dependent C_θ . So there is only a small shift in the position of the posterior's maxima, and the 1σ intervals get a bit reduced when considering a model-dependent C_θ .

However this effect is not systematic and depends on the particular realization $\hat{\xi}$. In other words, approximating C_θ as a constant C results in a modeling error both for the position of the posterior's maxima and for the size of the 1σ intervals, which depends on the particular realization $\hat{\xi}$. We quantified this modeling error using a lot of SDSS DR7-Full simulations

$$\hat{\xi} \sim \mathcal{N}(\xi_{\theta_0}, C_{\theta_0})$$

For each parameter, ω_m and $D_V(0.3)$, we studied the error in the position of the posterior's maximum and in the size of the 1σ interval. We found a mean modeling error in the position of the posterior's maxima approximately equal to 20% to 30% of the 1σ intervals. The error in the size of the 1σ intervals is smaller and is approximately equal to 10%.

Finally we did the same analysis for next-generation surveys, simply by dividing the covariance matrix by a

factor c , with $c = 2$ and $c = 4$

$$\hat{\xi} \sim \mathcal{N}\left(\xi_{\theta_0}, \frac{1}{c} C_{\theta_0}\right)$$

We also found a small mean modeling error on the position of the posterior's maxima and on the size of the 1σ intervals. As the survey gets larger this modeling error decreases. More precisely if we multiply the size of the SDSS DR7-Full survey by a factor 4, the mean modeling error on the position of the posterior's maxima reaches $\approx 20\%$ of the 1σ interval size and the mean absolute error of the 1σ interval size reaches $\approx 6\%$.

So our conclusion is that the modeling error due to the approximation of C_θ as a constant C is quite small. However for a safer analysis (though this modeling error cannot be handled for sure), one can multiply the size of 1σ intervals by a factor ≈ 1.3 .

This conclusion is a bit surprising since we found a strong dependence of C_θ on θ . However there is a competing effect at work in the likelihood $\mathcal{L}_\theta(\xi)$ that tends to erase scaling effects.

Computing C_θ with a higher dimensional-parameter θ seems very difficult and cannot be addressed with our procedure yet. The approach proposed in Xu et al. (2012) of a semi-analytic C_θ seems very promising in that respect. However it requires an ad hoc fitting of some parameters. In order to perform this parameter fitting, our simulations for a 3-dimensional parameter θ could be very interesting to use. Such an analysis would enable to see whether our conclusions are still correct when considering a full set of cosmological parameters.

Part of this work was supported by the European Research Council grant ERC-228261. We would like to thank the anonymous referee for helping to improve the quality of this paper.

Funding for the SDSS and SDSS-II has been provided by the Alfred P. Sloan Foundation, the Participating Institutions, the National Science Foundation, the U.S. Department of Energy, the National Aeronautics and Space Administration, the Japanese Monbukagakusho, the Max Planck Society, and the Higher Education Funding Council for England. The SDSS Web Site is <http://www.sdss.org/>.

APPENDIX

OPTIMAL LINEAR COMBINATION OF ESTIMATORS

In this section we assume that we have two independent and unbiased Gaussian estimators X_1, X_2 (of dimension n) of X_0 with respective covariance matrices C_1 and C_2

$$X_1 \sim \mathcal{N}(X_0, C_1) \tag{A1}$$

$$X_2 \sim \mathcal{N}(X_0, C_2) \tag{A2}$$

We consider an unbiased estimator X of X_0 as a linear combination of X_1 and X_2

$$X = AX_1 + (Id - A)X_2 \tag{A3}$$

with A a square $n \times n$ matrix. The resulting covariance matrix is given by

$$C = \mathbb{E}[XX^T] = AC_1A^T + (Id - A)C_2(Id - A)^T \tag{A4}$$

where we used the fact that X_1 and X_2 are independent. We will show that the following choice of A gives an

extremum of $\det(C)$

$$A = (C_1^{-1} + C_2^{-1})^{-1} C_1^{-1} \quad (\text{A5})$$

$$Id - A = (C_1^{-1} + C_2^{-1})^{-1} C_2^{-1} \quad (\text{A6})$$

For this we use the following derivatives formulae, with B a symmetric $n \times n$ matrix

$$\frac{\partial \det(C)}{\partial C} = \det(C) C^{-T} \quad (\text{A7})$$

$$\frac{\partial (ABA^T)}{\partial A_{ji}} = AB J^{ij} + J^{ji} B A^T = AB J^{ij} + (AB J^{ij})^T \quad (\text{A8})$$

with $(J^{ij})_{kl} = \delta_{ik} \delta_{jl}$. Differentiating $\det(C)$ with respect to A , we get

$$\frac{\partial \det(C)}{\partial A_{ji}} = \sum_{kl} \det(C) C_{kl}^{-T} \frac{\partial C_{kl}}{\partial A_{ji}} \quad (\text{A9})$$

So it is sufficient to have for all i, j that $\frac{\partial C}{\partial A_{ji}} = 0$

$$\frac{\partial C}{\partial A_{ji}} = \frac{\partial (AC_1 A^T)}{\partial A_{ji}} + \frac{\partial ((Id - A)C_2(Id - A)^T)}{\partial A_{ji}} \quad (\text{A10})$$

$$= (AC_1 - (Id - A)C_2)J^{ij} + [(AC_1 - (Id - A)C_2)J^{ij}]^T \quad (\text{A11})$$

Again it is sufficient to only have $AC_1 - (Id - A)C_2 = 0$, which gives

$$A(C_1 + C_2) = C_2 \quad (\text{A12})$$

$$A = C_2 (C_1 + C_2)^{-1} \quad (\text{A13})$$

$$A = C_2 C_2^{-1} (C_1^{-1} + C_2^{-1})^{-1} C_1^{-1} \quad (\text{A14})$$

So we obtain the solution given by equations (A6) and equations (A6)

$$A = (C_1^{-1} + C_2^{-1})^{-1} C_1^{-1} \quad (\text{A15})$$

$$Id - A = (C_1^{-1} + C_2^{-1})^{-1} C_2^{-1} \quad (\text{A16})$$

Finally when using this expression of A into equation (A4) we get

$$C = \mathbb{E}[XX^T] = (C_1^{-1} + C_2^{-1})^{-1} \quad (\text{A17})$$

REFERENCES

- Albrecht, A. et al. 2006, arXiv:astro-ph/0609591
Amanullah, R. et al. 2010, ApJ, 716, 712
Anderson, L. et al. 2012, arXiv:1203.6594
Arnalte-Mur, P., Labatie, A., Clerc, N., Martínez, V. J., Starck, J.-L., Lachièze-Rey, M., Saar, E., & Paredes, S., A&A, 542, A34
Bassett, B., & Hlozek, R. 2010, in Dark Energy: Observational and Theoretical Approach, ed. P. Ruiz-Lapuente (Cambridge University Press), 246
Beutler, F., Blake, C., Colless, M., Jones, D.H., Staveley-Smith, L., Campbell, L., Parker, Q., Saunders, W., & Watson, F. 2011, MNRAS, 416, 3017
Berlind, A.A. & Weinberg, D.H. 2002, ApJ, 575, 587
Blake, C. et al. 2011a, MNRAS, 415, 2892
Blake, C. et al. 2011b, MNRAS, 418, 1707
Cole, S. et al. 2005, MNRAS, 362, 505
Croce, M. & Scoccimarro, R. 2006, Phys. Rev. D, 73, 063519
Eisenstein, D.J., & Hu, W. 1998, ApJ, 496, 605
Eisenstein, D.J. et al. 2001, AJ, 122, 2267
Eisenstein, D.J. et al. 2005, ApJ, 633, 560
Eisenstein, D.J., Seo, H.-J., & White, M. 2007, ApJ, 664, 660
Hamilton, A.J.S. 1993, ApJ, 417, 19
Ho, S. et al. 2012, arXiv1201.2137
Kaiser, N. 1986, MNRAS, 222, 323
Kazin, E.A. et al. 2010, ApJ, 710, 1444
Komatsu, E. et al. 2009, ApJS, 180, 330
Labatie, A., Starck, J.-L., Lachièze-Rey, M., & Arnalte-Mur, P. 2012, Statistical Methodology, 9, 85
Labatie, A., Starck, J.-L., & Lachièze-Rey, M. 2012, ApJ, 746, 172
Landy, S.D. & Szalay, A.S. 1993, ApJ, 412, 64
Manera, M. et al. 2012, arXiv1203.6609
Martínez, V.J., Arnalte-Mur, P., Saar, E., de la Cruz, P., Pons-Bordería, M.J., Paredes, S., Fernández-Soto, A., & Tempel, E. 2009, ApJ, 696, L93
Mehta, K.T., Cuesta, A.J., Xu, X., Eisenstein, D.J., & Padmanabhan, N. 2012, arXiv1202.0092
Padmanabhan, N. et al. 2007, MNRAS, 378, 852
Padmanabhan, N., & White, M. 2008, Phys. Rev. D, 77, 123540
Padmanabhan, N., Xu, X., Eisenstein, D.J., Scalzo, R., Cuesta, A.J., Mehta, K.T., & Kazin, E. 2012, arXiv1202.0090
Percival, W.J., Cole, S., Eisenstein, D.J., Nichol, R.C., Peacock, J.A., Pope, A.C., & Szalay, A.S. 2007, MNRAS, 381, 1053
Percival, W.J. et al. 2010, MNRAS, 401, 2148
Perlmutter, S. et al. 1999, ApJ, 517, 565
Pons-Bordería, M.-J., Martínez, V.J., Stoyan, D., Stoyan, H., & Saar, E. 1999, ApJ, 523, 480
Press, W.H., Teukolsky, S.A., Vetterling, W.T., & Flannery, B.P. 2007, Numerical Recipes 3rd Edition: The Art of Scientific Computing (Cambridge University Press)
Refregier, A., Amara, A., Kitching, T.D., & Rassat, A. 2011, A&A, 528, A33
Reid, B.A. et al. 2010, MNRAS, 404, 60

- Riess, A.G et al. 1998, AJ, 116, 1009
- Sánchez, A.G., Baugh, C.M., & Angulo, R. 2008, MNRAS, 390, 1470
- Sánchez, A.G., Crocce, M., Cabré, A., Baugh, C.M., & Gaztañaga, E. 2009, MNRAS, 400, 1643
- Smith, R.E., Peacock, J.A., Jenkins, A., White, S.D.M., Frenk, C.S., Pearce, F.R., Thomas, P.A., Efstathiou, G., & Couchman, H.M.P. 2003, MNRAS, 341, 1311
- Tegmark, M. et al. 2006, Phys. Rev. D, 74, 123507
- Tian, H.J., Neyrinck, M.C., Budavári, T., & Szalay, A.S. 2011, ApJ, 728, 34
- White, M. et al. 2011, ApJ, 728, 126
- Xu, X., Padmanabhan, N., Eisenstein, D.J., Mehta, K.T., & Cuesta, A.J. 2012, arXiv:1202.0091